FEBS 29836 FEBS Letters 579 (2005) 4685–4691

CO: A chemical ontology for identification of functional groups and semantic comparison of small molecules

Howard J. Feldman[a], Michel Dumontier[a,1], Susan Ling[a], Norbert Haider[b], Christopher W.V. Hogue[a,c,*]

[a] The Blueprint Initiative of the Samuel Lunenfeld Research Institute, Mount Sinai Hospital, 600 University Ave., Toronto, ON, Canada M5G 1X5
[b] Department of Drug Synthesis, Faculty of Life Sciences, University of Vienna, Althanstraße 14, A-1090 Vienna, Austria
[c] Department of Biochemistry, University of Toronto, 1 King's College Circle, Toronto, ON, Canada M5S 1A8

Received 13 June 2005; revised 21 July 2005; accepted 21 July 2005

Available online 2 August 2005

Edited by Robert B. Russell

Abstract A novel chemical ontology based on chemical functional groups automatically, objectively assigned by a computer program, was developed to categorize small molecules. It has been applied to PubChem and the small molecule interaction database to demonstrate its utility as a basic pharmacophore search system. Molecules can be compared using a semantic similarity score based on functional group assignments rather than 3D shape, which succeeds in identifying small molecules known to bind a common binding site. This ontology will serve as a powerful tool for searching chemical databases and identifying key functional groups responsible for biological activities.

Keywords: Ontology; Small molecule; Functional group; Pharmacophore; Semantic similarity

1. Introduction

Small molecules play a crucial role in the modulation of biological function, not to mention serving as metabolites for building blocks of larger biopolymers such as DNA and protein. Protein–small-molecule interactions are captured by a variety of experimental methods. Those determined by X-ray crystallography are deposited in the protein data bank (PDB) structure database [1]. The 3DSM division of the Biomolecular Interaction Network Database (BIND) catalogues 22 367 non-redundant protein–small molecule interactions while taking steps to remove spurious interactions such as interactions with ions that do not likely fulfil a biological role [2,3]. The Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database places metabolites, compounds and drugs in biochemical reactions and pathways, mapping reference biological pathways to many organisms [4]. These databases have a limited number of small molecules, but other databases such as ZINC [5], the developmental therapeutics program (DTP) [6] at NCI, Chembank (http://chembank.broad.harvard.edu/), and PubChem at NCBI (http://pubchem.ncbi.nlm.nih.gov/) have increased the number of readily available small molecules to over one million. In fact, PubChem is a resource that intends to be a comprehensive repository for chemical structures of small organic molecules along with information on their biological activities. This increase in publicly available small molecules will drive new efforts to better understand interactions involving small-molecules, particularly in the area of drug docking and pharmacogenomics. However, a significant challenge exists to identify the important underlying sets of functional groups of small molecules involved in biological interactions, or 'pharmacophores', and to use this information to recognize other, possibly more biologically active small molecules.

Ontologies, or controlled vocabularies, have been shown to be extremely useful to researchers, in order to aid in the classification and organization of information. For example, the successful gene ontology (GO) [7] is used to describe myriad information regarding protein function and classification, to a fine level of detail. Relationships within an ontology can be used to help group together similar objects, or find things with similar properties or behaviours. The majority of ontologies are used to standardize definitions, as well as facilitate data exchange, analysis, and searching. Applying a suitable small-molecule ontology to a list of small-molecule interactions would allow users to search for chemicals possessing functional groups complementary to those found in the binding pocket of a protein of interest.

To our knowledge, no attempts have been made to arrive at a formal chemical ontology for describing small molecules until recently. The sole exception is chemical entities of biological interest (ChEBI) [8], a small-molecule database hosted at the European Bioinformatics Institute (EBI), which has developed an ontology to help classify small molecules in their database. Each entry in the database is manually added as a leaf along one or more branches in the ontological tree, with approximately 11 000 total terms (including leaf terms). While simple classifications such as aldehydes, ketones, alcohols, etc., naturally lend themselves as ontological terms for chemicals, ChEBI goes into much more detail than this, with many levels of specificity within

*Corresponding author. Fax: +1 416 596 8077.
E-mail address: chogue@blueprint.org (C.W.V. Hogue).

[1] Present address: Department of Biology, Carleton University, 1125 Colonel By Drive, Ottawa, ON, Canada K1S 5B6.

Abbreviations: SMID, small molecule interaction database; PDB, Protein Data Bank; ChEBI, chemical entities of biological interest; GO, gene ontology; ACPC, 1-aminocyclopropane-1-carboxylic acid; CO, chemical ontology

0014-5793/$30.00 © 2005 Published by Elsevier B.V. on behalf of the Federation of European Biochemical Societies.
doi:10.1016/j.febslet.2005.07.039

the ontological tree, and grouping both by chemistry and by molecular function. However, each time a new compound is added to the ChEBI database, it must be assigned to the ontological tree by hand, which is very labour-intensive and somewhat subjective. Many of the terms themselves are somewhat vague and open to interpretation, since no strict definitions of them are supplied. Because ChEBI terms may have multiple parents, it will become gradually more difficult to establish and maintain relationships in a growing ontology, as was found for classification of terms in the medical domain [9]. Multiple relationships are better handled by formal expressiveness and the reasoning capabilities of an underlying description logic such as DAML + OIL, hence the development of a new methodology for GO to increase its formal explicit semantic content [10]. Since small molecules can be considered as comprising largely independent chemical functional groups, a classification for small molecules could also be obtained by identifying the molecule's chemical functional groups using objective and computable criteria. A strictly hierarchical approach to small molecule classification is desirable in maintaining simplicity, allowing simple searching and ontological assignment while providing a set of rich descriptive terms that can be used for semantic comparison.

We present a new small molecule chemical ontology (CO) based on functional groups assigned by the program checkmol (http://merian.pch.univie.ac.at/~nhaider/cheminf/cmmm.html). The major advantages it offers are that: (a) the terms can be automatically, consistently and objectively assigned by a computer program; (b) as a result precise definitions are available for each term, with most including a chemical sketch of the functional group and (c) its simplicity facilitates computational applications.

Often drug molecules can be enhanced in potency through minor changes to their chemical structure to allow tighter binding, or better packing in the binding site. Similar molecules tend to bind the same pocket, albeit with different affinities. In fact, it is well known that transition state analogues of enzyme substrates act as ideal inhibitors of those same enzymes [11]. The portion of a small molecule responsible for molecular recognition, or binding, in a particular binding site is referred to as a 'pharmacophore'. This could be described by numbers of hydrogen bond donors/acceptors in the binding site, charges if any, aromatic ring centres, and so forth. Thus, it is not necessarily desirable to compare molecules based on size or shape, but rather based on chemical properties. A pharmacophore-based approach to similarity seems logical if the purpose is to find more potential targets for a given binding site with known small molecule ligands. Semantic similarity measures have been applied previously to the GO and were shown to have some correlation with sequence similarity [12]. Hence, as a demonstration of the utility of CO, we employ it to generate a semantic similarity metric, similar to the Tanimoto score [13] but making use of the ontology, which is applied across the PubChem database and is able to identify similar molecules based on functional groups alone. Additionally, it can be used as a powerful search interface, allowing one to pull out structures with any desired combination of functional groups. It has also been integrated with the Small Molecule Interaction Database (SMID), which is used to detect underlying pharmacophores in sets of predicted protein–small molecule interactions.

2. Materials and methods

2.1. Automatic detection of functional groups

Checkmol is a freely available, open-source tool which is able to automatically detect and assign functional group information to any small molecule with 2D co-ordinates. It can identify over 200 functional groups such as 'secondary carboxylic acid amide' or 'sulfonic acid derivative'. Assignment is made strictly based on computational detection of specific arrangements of atoms and bonds within the molecule, and thus is completely objective and fully automated. A given structure may have any number of functional groups assigned to it, but each term will only appear once, if multiple instances of the functional group appear in the molecule. An interactive version of the program is available at: http://merian.pch.univie.ac.at/~nhaider/fga.php. When run on the command line, a standard mol file is given as input, and the functional groups are output, one per line.

2.2. CO – The chemical ontology

We have studied the checkmol functional group terms and their relationships to one another, and used this to come up with a new ontology (Fig. 1). In order to allow grouping of similar terms together, it was necessary to add a few terms of our own for a total of 231 distinct terms. In this ontology, each term has precisely one parent. The complete ontology is available for download at ftp://ftp.blueprint.org/pub/SMID/ontology/CO.obo. Only the most specific terms are assigned to molecules. Hence, a molecule may have two sibling terms assigned, but not a child and parent, or child and grandparent, to the same molecule. This simply helps to remove redundancy in the assignments, since any child term automatically implies the parent terms as well.

2.3. A metric for semantic similarity

To begin, all the checkmol functional group assignments are recorded for each of the two molecules being compared (discarding parents as noted above). Additionally, we determine the frequency of each term in the small-molecule database, i.e., how many molecules in the database each term applies to. Denote the frequency of a term T by f_T. Then, let S_T denote the score assigned when two terms match in different molecules, where $S_T = f_T^{-1/4}$. Terms which occur frequently will be down-weighted, while terms which have a low frequency, and yet are found to be common to two molecules, will contribute a score close to 1. The exponent 1/4 was chosen to result in a score of about 0.1 for common terms, compared to close to 1 for rarely assigned ones, for a database size on the order of 10 000 molecules.

Next, we define two terms as siblings if their immediate parent is the same, or if one term is the immediate parent of the other. Note that because all redundant parents have been removed, the presence of a non-leaf node A of the ontology implies that there exists some child D of A without a term yet assigned to it (see Fig. 2). Hence, A represents a missing term D in the ontology which is a direct child of A, and it can be considered a sibling to any other children B, C of A. For example, if 'acetal' were assigned (A), this means that it must be an acetal compound other than the current children: carbonyl hydrate (B) or hemiacetal (C). Whatever specifically it may be, it can be considered a sibling to a hemiacetal, for example, since it is really some specific type of acetal compound which we have just not defined a term for yet (D). The 'pairwise score', z, for two small molecules is obtained by matching their terms:

$$z = \sum_{\text{matched terms } T} S_T + \sum_{\text{siblings } T1,\ T2} \frac{1}{\frac{1}{S_{T1}} + \frac{1}{S_{T2}}} \tag{1}$$

Before computing sibling matches, all exactly matching terms are removed from the term lists of each molecule to avoid double counting.

The pairwise score is then normalized to account for the different number of terms assigned to each molecule. If the two molecules being compared have a total of x and y ontological terms assigned to them, respectively, then the final similarity score is obtained from:

$$\text{score} = \frac{2z}{x+y} \tag{2}$$

This is similar to the Tanimoto score [13], and results in a score between zero and one with a higher score indicating more functional groups in common. The maximum value varies with the frequency of occurrence of the functional group terms. We also compute a score

1,2-alcohol derivative [30054]
  1,2-aminoalcohol [8288]
  1,2-diol [23262]
alkyne [7368]
amine [168969]
  primary amine [58308]
    primary aliphatic amine (alkylamine) [22569]
    primary aromatic amine [36540]
  secondary amine [44631]
    secondary aliphatic amine (dialkylamine) [22018]
    secondary aliphatic/aromatic amine (alkylarylamine) [16252]
    secondary aromatic amine (diarylamine) [7516]
  tertiary amine [83570]
    tertiary aliphatic amine (trialkylamine) [64581]
    tertiary aliphatic/aromatic amine (alkylarylamine) [23737]
    tertiary aromatic amine (triarylamine) [116]
anion [36860]
aromatic compound [467245]
  (het)arene derivative [66786]
    imino(het)arene [15809]
    oxo(het)arene [52744]
    thioxo(het)arene [1155]
  1,2-diphenol [3430]
  phenol or hydroxyhetarene [51090]
azide [2519]
azo compound [10972]
bis-acyl compounds [24528]
  carboxylic acid anhydride [1171]
  carboxylic acid imide [23380]
    carboxylic acid imide, N-substituted [11621]
    carboxylic acid imide, N-unsubstituted [12285]
boronic acid derivative [1044]
  boronic acid [170]
  boronic acid ester [765]
carbodiimide [70]
carboxylic/carbonic acid sp2 derivative [372933]
  alpha-aminoacid [5088]
  alpha-hydroxyacid [1963]
  carbonic acid dervivative [56886]
    carbamic acid derivative [23618]
      carbamic acid [882]
      carbamic acid ester (urethane) [22738]
      carbamic acid halide (haloformic acid amide) [43]
    carbonic acid diester [1025]
    carbonic acid ester halide (alkyl/aryl haloformate) [79]
    carbonic acid monoester [247]
    semicarbazide [936]
    urea [31868]
  carboxyl/carbonyl group [344947]
    acyl cyanide [23]
    carbonyl derivative [96130]
      carbonyl compound [95911]
        aldehyde [7900]
        ketone [78166]
      thiocarbonyl compound [257]
        thioaldehyde [67]
        thioketone [142]
    carboxylic acid derivative [283665]
      acyl halide [717]
        acyl bromide [23]
        acyl chloride [619]
        acyl fluoride [73]
        acyl iodide [2]
      carboxylic acid [58690]
      carboxylic acid amide [101101]
        primary carboxylic acid amide [11081]
        secondary carboxylic acid amide [73386]
        tertiary carboxylic acid amide [29034]
      carboxylic acid azide [36]
      carboxylic acid ester [101449]
      carboxylic acid hydrazide [5638]
      carboxylic acid salt [5130]
      hydroxamic acid [1221]
    thiocarboxylic acid derivative [3147]
      thiocarboxylic acid [169]
      thiocarboxylic acid amide [440]
      thiocarboxylic acid ester [2105]
  thiocarbonic acid derivative [6298]
    thiocarbamic acid derivative [4199]
      thiocarbamic acid [805]
      thiocarbamic acid ester [3391]
      thiocarbamic acid halide (halothioformic acid amide) [4]
    thiocarbonic acid diester [536]
    thiocarbonic acid ester halide (alkyl/aryl halothioformate) [7]
    thiocarbonic acid monoester [299]
    thiosemicarbazide [65]
    thiourea [1209]
carboxylic acid amide derivative [6406]
  carboxylic acid amidine [6065]
  carboxylic acid amidrazone [341]
cation [47111]
CN sp derivative [20092]
  carbonitrile [19628]
  cyanate [24]
  thiocyanate [441]
CO2/carboxylic acid sp3 derivative [63273]
  carboxylic acid sp3 derivative [62917]
    acetal/aminal derivative [61686]
      acetal [29462]
        carbonyl hydrate [372]
        hemiacetal [3127]
      aminal [6687]
      hemiaminal [21738]
      hemithioaminal [4358]
      thioacetal [1297]
    orthocarboxylic acid derivative [1404]
      amide acetal [102]
      orthoester [407]
  CO2 derivative (general) [359]
diazonium salt [642]
dioxide derivative [71244]
  disulfide [3394]
  hydrazine derivative [31427]
  hydroperoxide [480]
  hydroxylamine [38940]
  peroxide [557]
ene derivative [124526]
  alkene [109073]
  enol/enamine derivative [25295]
    enamine [16821]
    enol derivative [9789]
      enol [3877]
      enol ether [6605]
enediol [75]
ether derivative (general) [274890]
  ether derivative [149227]
    ether [128019]
      alkyl aryl ether [87162]
      dialkyl ether [41268]
      diaryl ether [6052]
    thioether [24307]
  hydroxy compound [145089]
    alcohol [103753]
      primary alcohol [41311]
      secondary alcohol [67762]
      tertiary alcohol [20109]
  sulfenic acid derivative [17899]
    sulfenic acid [7]
    sulfenic acid amide [156]
    sulfenic acid ester [35]
    sulfenic acid halide [17]
  thiol (sulfanyl compound) [7961]
    alkylthiol [3291]
    arylthiol [4678]
halogen derivative [147358]
  alkyl halide [33014]
    alkyl bromide [7048]
    alkyl chloride [18583]
    alkyl fluoride [6926]
    alkyl iodide [1355]
  aryl halide [98918]
    aryl bromide [16186]
    aryl chloride [63898]
    aryl fluoride [19280]
    aryl iodide [4704]
heterocyclic compound [352104]
  lactam [14196]
  lactone [12818]
  thiolactam [93]
  thiolactone [189]
imine derivative [13933]
  imido ester [1284]
  imidothioester [702]
  imidoyl halide [349]
  imine [11632]
isocyanate derivative [1194]
  isocyanate [514]
  isothiocyanate [681]
isonitrile [203]
isourea derivative [18344]
  guanidine [8029]
  isothiourea [9443]
  isourea [1013]
ketene [44]
ketene acetal or derivative [4247]
N-oxide [38365]
nitro derivative [34223]
  nitrate [252]
  nitro compound [34000]
nitroso derivative [1795]
  nitrite [160]
  nitroso compound [1635]
organometallic compound [3417]
  organolithium compound [0]
  organomagnesium compound [0]
oxime/-zone derivative [12302]
  hydrazone [5879]
  oxime [2608]
  oxime ether [2560]
  semicarbazone [1231]
  thiosemicarbazone [33]
phosphine [1486]
phosphoric/phosphonic acid group [19120]
  phosphine oxide [582]
  phosphonic acid derivative [5241]
    phosphonic acid [1725]
    phosphonic acid ester [2944]
  phosphoric/thiophosphoric acid derivative [13467]
    phosphoric acid derivative [11973]
      phosphoric acid [491]
      phosphoric acid amide [2603]
      phosphoric acid ester [10294]
      phosphoric acid halide [226]
    thiophosphoric acid derivative [1553]
      thiophosphoric acid [2]
      thiophosphoric acid amide [427]
      thiophosphoric acid ester [1375]
      thiophosphoric acid halide [82]
quaternary ammonium salt [11226]
sulfinic acid group [2342]
  sulfinic acid derivative [629]
    sulfinic acid [177]
    sulfinic acid amide [77]
    sulfinic acid ester [75]
    sulfinic acid halide [2]
  sulfoxide [1713]
sulfuric/sulfonic acid derivative [45936]
  sulfone [7344]
  sulfonic acid derivative [33934]
    sulfonamide [17639]
    sulfonic acid [11735]
    sulfonic acid ester [3528]
    sulfonyl halide [1588]
  sulfuric acid derivative [5605]
    sulfuric acid [1738]
    sulfuric acid amide [314]
    sulfuric acid amide ester [222]
    sulfuric acid diamide [572]
    sulfuric acid diester [72]
    sulfuric acid monoester [2053]
    sulfuryl halide [61]

Fig. 1. The complete Chemical Ontology applied to the PubChem database. Number of molecules (out of 636359 total) falling under each category is given in brackets. Each parent node includes counts for all the children nodes as well, so for example there are 168969 molecules with at least one amine group, of which 58308 have at least one primary amine, 44631 secondary and 83570 tertiary. Some molecules have more than one subtype, for example primary and tertiary, in the same molecule, so that these three counts add up to more than the total number of molecules with amines.

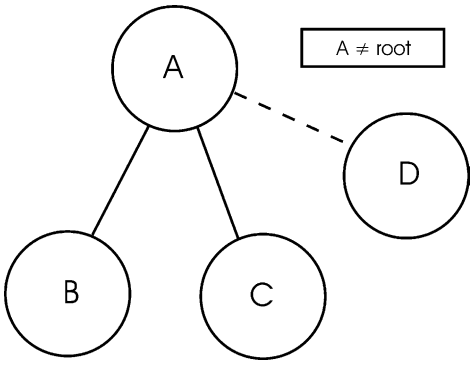


Fig. 2. Parent node A has two known children, B and C; any members of A which are not B or C are lumped into imaginary child D. See text for an example.

with $S_T = 1$ for all T, termed the 'non-weighted score'. We add a further restriction that if the ratio of molecular weights of two molecules is greater than 3 (or less than 1/3), they automatically receive a similarity score of zero. By looking at similarity hits to several popular compounds such as ATP and benzene, it was empirically determined that using a non-weighted score of 0.6 as the cutoff produced reasonable numbers of hits, at reasonable levels of similarity to the query. The non-weighted score must be used for the cutoff as it always ranges between zero and one, while the normal similarity score peak value varies with the functional groups involved.

As an example, let us say we have two molecules A and B in our database of 20 000 that we wish to compare. A is identified as having the following four functional groups, with frequency in the database given in parentheses: sulfuric acid monoester (117), nitrate (5), sulfuric acid diamide (16) and acetal (1856). B has only three functional groups: sulfuric acid diester (2), sulfuric acid diamide (16) and hemiacetal (465). Thus, A and B have one functional group in common, sulfuric acid diamide and $S_{\text{sulfuric acid diamide}} = 16^{-1/4} = 0.5$. The terms sulfuric acid monoester and sulfuric acid diester are siblings in the ontology, with $S_{\text{sulfuric acid monoester}} = 0.304$ and $S_{\text{sulfuric acid diester}} = 0.841$. Note that these are both siblings to sulfuric acid diamide as well, but we have already computed this term's contribution to the score in the first part of the calculation. Lastly, note that acetal is a parent of hemiacetal, and so these terms are treated as siblings as explained above. $S_{\text{acetal}} = 0.152$ and $S_{\text{hemiacetal}} = 0.215$. Note that the frequency of acetal in the database, 1856 occurences, includes only those where a more specific child was not assigned. Now we can compute the score. First for the non-weighted score, $S_T = 1$ for all terms T, so in Eq. (2), $x = 4$, $y = 3$ and $z = (1 + 0.5 + 0.5) = 2$, giving a result of $2*2/(4 + 3) = 0.57$. The frequency weighted score is then computed from Eq. (1) as $z = 0.5 + 0.223 + 0.089 = 0.812$, thus from Eq. (2), score $= 2*0.812/(4 + 3) = 0.232$. Note that as expected, the sibling terms contribute less to the score than the matching terms, and the more common acetal siblings contribute only 0.089 compared with the rarer sulfuric acid derivative siblings which contribute 0.223.

3. Results and discussion

3.1. Use of semantic similarity

For a given small molecule, the similarity score can be used to find 'neighbours', or similar molecules, in the database. The top hits will generally have most of the same functional groups as the target molecule, with perhaps a few extra. Molecules with similar functional groups but with many additional ones will generally rank lower. Molecules which share functional groups which are very commonly assigned in the database but not rare ones will also tend to score lower than those which share at least one rarely assigned term. The molecular weight cutoff helps to avoid marking grossly different sized molecules as similar, even if they share similar functional groups to some degree. Fig. 3 shows penicillin G and its top four neighbours. All are penicillin-like and clearly similar to the query as would be expected. In the top 20 hits, only thiazoloisoindolinone – an HIV-1 reverse transcriptase inhibitor and thiazolidine are not penicillin-like.

The similarity score should also be able to group molecules that bind a common site competitively. Ideally, any one of these ligands of the protein should receive a high rank when compared to any of the others that bind the same site. Unfortunately, there are no complete lists of all small molecules that bind a particular site (within some threshold affinity constant), therefore we are unable to fully quantify the accuracy of the similarity score at this time. However, there are certain experimentally demonstrated examples which we can use to validate the chemical neighbouring capability of this approach. We chose one such example, comparing several ligands of a protein, all known to bind the same binding site and already in PubChem. These were a number of tetra-peptides which act as potent protease inhibitors of Hepatitis C Virus NS3 [14]. Compounds 9, 16 and 25 in Johansson et al.'s work correspond to PubChem CIDs 497549, 497564, and 497573, respectively.

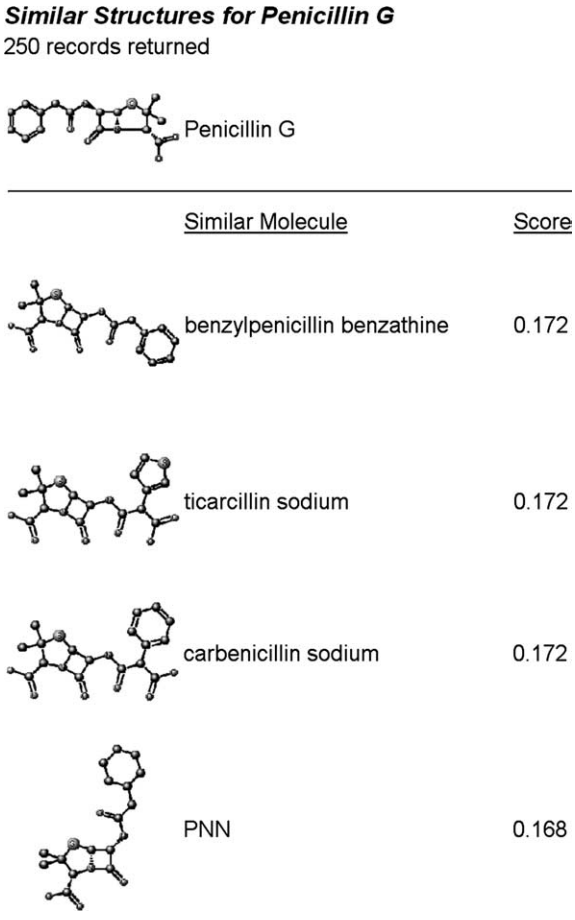


Fig. 3. Penicillin G and its top four hits by similarity score, in our local database of approximately 20 000 small molecules. Next to the image is the molecule name, followed by the similarity score. PNN is the same as the query but with the aromatic ring incorrectly drawn as all single bonds (this will be corrected by automatic bond order assignment in the near future).

The former two are the most potent inhibitors of the 25 tested by them, while the latter is the most potent one containing 1-aminocyclopropane-1-carboxylic acid (ACPC). Previous structure–activity relationship studies had shown that Cys or ACPC must be in the P1 position of the peptide. First, CID:497549 was compared to all of PubChem (636 359 structures), with 7221 compounds receiving significant matches (non-weighted score above 0.6). The resulting scores and ranks are summarized in Table 1. CID:497564 ranks highest out of all the compounds, while CID:497573 ranks 971st. The PubChem website offers 26 similar structures with CID:497564 ranking 13th using the same structure-similarity method as the NCI database, from the CACTVS toolkit [15]. Next, using CID:497564 as the query, 7146 hits are returned, with CID:497549 ranking first and CID:497573 ranking 965th. CID:497549 ranks 30th out of 51 neighbours on the PubChem web site. Lastly, comparing CID:497573 to PubChem gives 4274 hits back with the other two molecules tied at 1744th. Neither of these two appear among the 14 neighbours at the PubChem website. Other high-scoring hits to the first two small molecules involved amino-carboxylic acid derivatives, benzoic acid derivatives and peptides, all with a sulfur moiety or Cys residue. Because CID:497573 lacks the Cys group, this pushed down its score when compared to the other two. It scored better against amino-carboxylic acids and peptides without a sulfur moiety in the molecule.

Table 1
Similarity scores between three tetra-peptides all known to bind the same binding site of Hepatitis C Virus NS3

Hits	Query		
	CID:497549	CID:497564	CID:497573
CID:497549		0.0744 (1)	0.0416 (1744)
CID:497564	0.0744 (1)		0.0416 (1744)
CID:497573	0.0416 (971)	0.0416 (965)	

Compounds are identified by their PubChem Compound IDentifiers (CID). A higher score is better. The rank of the score compared to all other molecules in PubChem is given in parentheses after each score.

This simple example demonstrates that our scoring system is able to pull out similar molecules, based on the requirement of having functional groups that may affect binding to the same binding site, in a comparable fashion to PubChem's similar compounds link, which uses the Cactvs

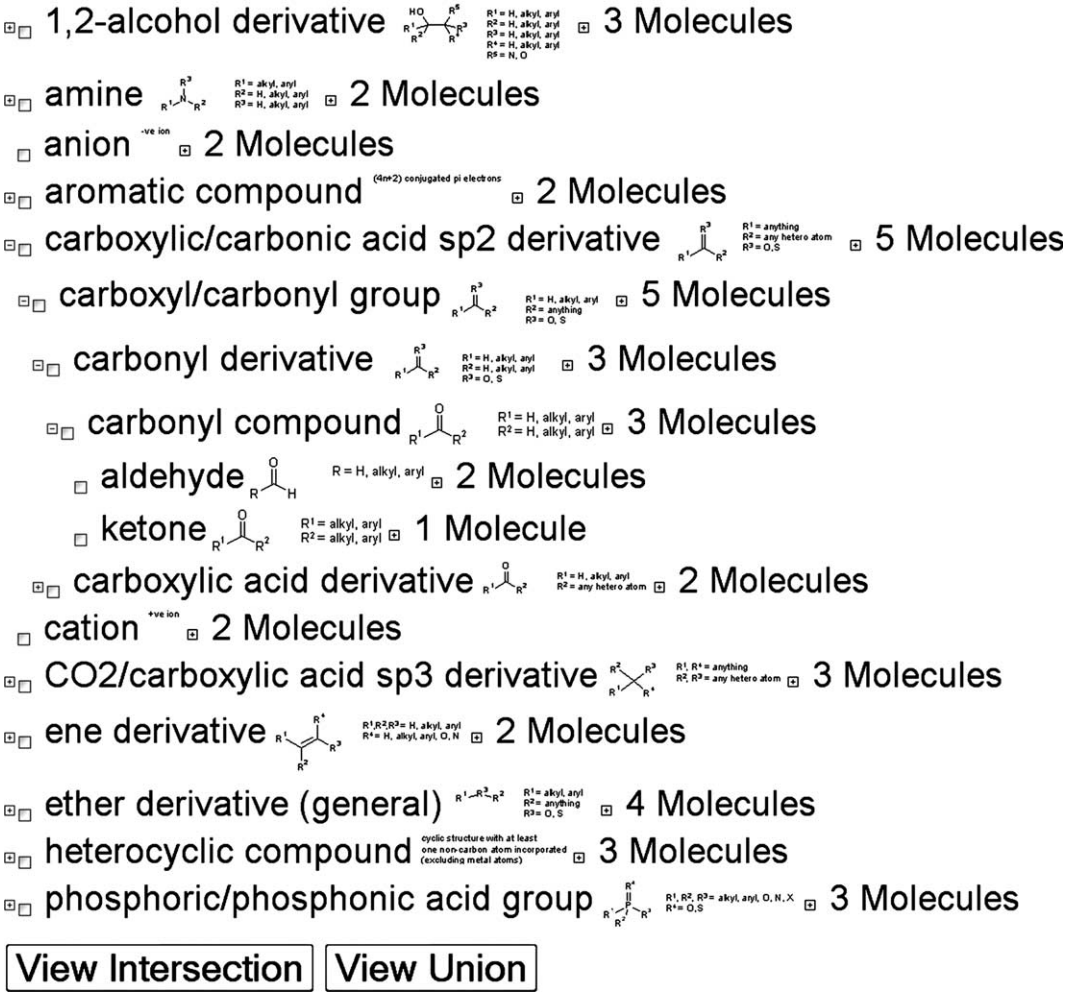
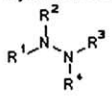


Fig. 4. Chemical ontology view for small molecules predicted to associate with primary binding site for *Escherichia coli* aldehyde dehydrogenase B (GI 38704182). The largest number of molecules (5) is associated with the carboxyl/carbonyl group, which indeed contains the substrate, aldehyde, and the product, carboxylic acid.

substructure-key ''fingerprint'', consisting of about 900 substructures, to locate similar structures in the database with a Tanimoto score of 90% or better. Unlike the standard Tanimoto score, the method described herein also accounts for functional group siblings, using the ontology tree, and accounts for the frequency of the terms in the database. However, returning 1% of the database in the hit list is too large to be of practical use in most cases. As more terms are added to the ontology, the scoring function will become more sensitive to the subtle differences between the different molecules, and ultimately return fewer hits, ranked more specifically on chemical groups added to the ontology that are likely to be important for binding.

Small molecules sharing functional groups:

- cation
  +ve ion
- hydrazine derivative
  

  R1 = H, acyl, alkyl, aryl
  R2 = H, acyl, alkyl, aryl
  R3 = H, acyl, alkyl, aryl
  R4 = H, acyl, alkyl, aryl
- dialkyl ether
  

  R1 = alkyl
  R2 = alkyl

64 matching molecules

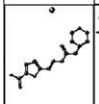
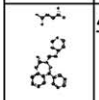
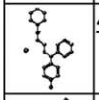
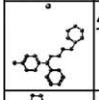
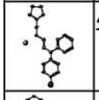
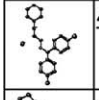
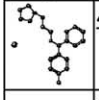
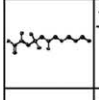
	Compound ID	IUPAC
	14164	N-[(5-nitro-2-furyl) methylideneamino]-2-(1-oxa-4-azoniacyclohex-4-yl)ethanamide chloride
	43072	4-[2-(1-oxa-4-azoniacyclohex-4-yl) ethyl]-6-phenyl-4,5-diazabicyclo[5.4.0]undeca-5,7,9,11-tetraen-3-one; 4-hydroxy-4-oxo-but-2-enoate
	43074	4-[3-(1-oxa-4-azoniacyclohex-4-yl) propyl]-6-phenyl-4,5-diazabicyclo [5.4.0]undeca-5,7,9,11-tetraen-3-one; 4-hydroxy-4-oxo-but-2-enoate
	45395	2-[(4-methylphenyl)-phenyl-methoxy] ethyl-morpholin-4-yl-ammonium chloride
	45387	2-benzhydryloxyethyl-morpholin-4-yl-ammonium chloride
	45397	N-[2-[(4-methylphenyl)-phenyl-methoxy]ethyl]-3,4,5,6-tetrahydro-2H-pyridin-1-amine chloride
	45328	2-[(4-bromophenyl)-phenyl-methoxy] ethyl-pyrrolidin-1-yl-ammonium chloride
	45617	2-[bis(4-chlorophenyl)methoxy] ethyl-morpholin-4-yl-ammonium chloride
	45399	N-[2-[(4-methylphenyl)-phenyl-methoxy]ethyl]-2,3,4,5-tetrahydropyrrol-1-amine chloride
	104167	1-[(3-butoxy-2-hydroxy-propyl)-dimethyl-ammonio]imino-2-methyl-prop-2-en-1-olate
	104168	(3-butoxy-2-hydroxy-propyl)-methacryloylamino-dimethyl-ammonium

Fig. 5. Molecules in PubChem containing all three functional groups that were queried: cation, hydrazine derivative, and dialkyl ether. 64 molecules in the database match these criteria. On the web-page, clicking a compound ID links back to the original PubChem record, while clicking on an image enlarges it.

3.2. Applying the chemical ontology

SMID is a database of domain-small molecule interactions involving only 4283 small molecules found in 3D structures. SMID provides a bridge between structure space and sequence space for small molecule binding annotation. The ontology has been applied to SMID, to further annotate the set of small molecules predicted to bind to a particular binding site of a protein. This set is viewed on an interactive ontological tree diagram as the query set is returned by the web interface. Only the terms applicable to the molecules in question (and their parents) will be displayed, for the user to navigate along. This permits one to quickly identify which features are common to all the small molecules in the binding site, and help identify the binding mode and even function at a glance. Fig. 4 illustrates the distribution of functional groups for the five small molecules predicted to bind the primary binding site of *Escherichia coli* O157:H7 aldehyde dehydrogenase B. The 'carboxyl/carbonyl group', which includes the aldehyde group, the substrate for this enzyme, as well as the product, carboxylic acid, was the most frequent functional group, assigned to all five of the different molecules. Thus, the ontology may aid in rapid identification of functional groups important in protein–small molecule interactions.

The ontology has been applied to a local copy of NCBI's PubChem database as well, allowing browsing of the database by functional group. The distribution of the molecules in PubChem across CO is shown in Fig. 1, and is available through an interactive website at http://smid.blueprint.org/pubchem/. The user may select several functional groups of interest and then view the intersection (or union) of these. This acts as a simple yet powerful query engine, performing filtering by functional group. For example, suppose prior research had indicated that desirable features of an inhibitor were that it contain a positive charge, a dialkyl ether group, and was a hydrazine derivative. Computing the intersection gives us a short list of 64 molecules matching the criteria (Fig. 5). There is no other way to perform such a complex query in such a simple intuitive manner.

We plan to extend the checkmol program to identify more biology-specific functional groups such as those for nucleotides, pyrimidines, purines, amino acids, saccharides, steroids and so forth. These would then be added to CO as well, and help to further categorize the larger clusters of molecules in the current tree to reduce the number of molecules falling into any one leaf node, and result in a much more powerful and sensitive similarity measure. This work offers only a brief glimpse at the possibilities made available through the use of a robust chemical ontology joined together with an appropriate objective ontology assignment tool. As CO grows, it will serve as a powerful tool for searching large chemical databases and identifying key functional groups responsible for biological activities.

Acknowledgements: Funding for this project has been provided from Genome Canada through the Ontario Genomics Institute and from the Ontario Research and Development Challenge Fund in grants to C.W.V.H.

References

[1] Berman, H.M. et al. (2002) The Protein Data Bank. Acta Crystallogr. D: Biol. Crystallogr. 58, 899–907.
[2] Salama, J.J., Donaldson, I. and Hogue, C.W. (2001) Automatic annotation of BIND molecular interactions from three-dimensional structures. Biopolymers 61, 111–120.
[3] Alfarano, C. et al. (2005) The Biomolecular Interaction Network Database and related tools 2005 update. Nucleic Acids Res. 33, D418–D424.
[4] Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. and Hattori, M. (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res. 32, D277–D280.
[5] Irwin, J.J. and Shoichet, B.K. (2005) ZINC – a free database of commercially available compounds for virtual screening. J. Chem. Inf. Model 45, 177–182.
[6] Monga, M. and Sausville, E.A. (2002) Developmental therapeutics program at the NCI: molecular target and drug discovery process. Leukemia 16, 520–526.
[7] Ashburner, M. et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25, 25–29.
[8] Brooksbank, C., Cameron, G. and Thornton, J. (2005) The European Bioinformatics Institute's data resources: towards systems biology. Nucleic Acids Res. 33, D46–D53.
[9] Rogers, J.E., Price, C., Rector, A.L., Solomon, W.D., Smejko, N., 1998. Validating clinical terminology structures: integration and cross-validation of Read Thesaurus and GALEN. Proc AMIA Symp, 845–849.
[10] Wroe, C.J., Stevens, R., Goble, C.A., Ashburner, M., 2003. A methodology to migrate the gene ontology to a description logic environment using DAML + OIL. Pac. Symp. Biocomput., 624–635..
[11] Schramm, V.L. (1998) Enzymatic transition states and transition state analog design. Annu. Rev. Biochem. 67, 693–720.
[12] Lord, P.W., Stevens, R.D., Brass, A. and Goble, C.A. (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics 19, 1275–1283.
[13] Resnik, P. (1999) Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. J. Art. Int. Res. 11, 95–130.
[14] Johansson, A., Poliakov, A., Akerblom, E., Lindeberg, G., Winiwarter, S., Samuelsson, B., Danielson, U.H. and Hallberg, A. (2002) Tetrapeptides as potent protease inhibitors of Hepatitis C Virus full-length NS3 (protease helicase/ NTPase). Bioorg. Med. Chem. 10, 3915–3922.
[15] Ihlenfeldt, W.D., Voigt, J.H., Bienfait, B., Oellien, F. and Nicklaus, M.C. (2002) Enhanced CACTVS browser of the Open NCI Database. J. Chem. Inf. Comput. Sci. 42, 46–57.